Project Thesaurus 2020 - Linguistic and Ontological Aspects

Sylvie Reusch [1], Peter Reusch [2], Michael Meschede [3]

[1] Université de Bourgogne, Dijon - sylviereusch@gmx.de
[2] Fachhochschule Dortmund - University of Applied Sciences and Arts - peter.reusch@fh-dortmund.de
[3] Fachhochschule Dortmund - University of Applied Sciences and Arts - michael.meschede@gmx.de

***Abstract*—Structures and linguistic concepts of thesauri are analyzed and compared. Proposals for the improvement of thesauri are developed.**

Keywords — thesaurus; linguistic concepts; analogy; ontologies; topic maps

I. INTRODUCTION

Thesauri are dictionaries with preferred terms in a hierarchy and several relations used to link synonyms and otherwise related terms. Thesauri are used to learn languages, to classify documents, to support creativity, and for many more reasons. Thesauri are available all over the world. Peter Mark Roget's "Thesaurus of English words and phrases" (1852) [1] has a new edition every year since more than 150 years - it´s a book like the bible - and it was a prototype for many other thesauri.

But today there is a crisis on thesauri. Regarding Roget´s Thesaurus, even the publishers have some problems to decide what they want. There are editions of Roget´s Thesaurus with the hierarchy of preferred terms and others without this hierarchy - and in this case the document is just a dictionary with words in an alphabetical order - not appropriate for concept learning and not appropriate to support creativity.

Roget's Thesaurus was adopted for other languages, for example, German [20], Spanish [19] and French [2]. In France, "Le dictionnaire idéologique" (1858) of Théodore Robertson is the first application of the thesaurus in French [2].

In 1992 the publisher Larousse decided to create a *new* French thesaurus: The Thésaurus Larousse [3]. The Thésaurus Larousse was inspired by Roget's Thesaurus, but is much weaker. Today the strong tradition of French publishers in alphabetically ordered dictionaries undermines the publication of a new paper version of Thésaurus Larousse and the development of other new types of linguistic thesauri in France - new thesauri with language updates and further developed semantic relations. Instead of developing a really new thesaurus Larousse tries to integrate the existing thesaurus into a more flexible Ipad® application. But without a new understanding of core thesaurus concepts that will not be so strong.

Beside the kind of thesauri mentioned so far there are many important non-linguistic thesauri. In contrast to linguistic thesauri these thesauri often contribute to knowledge management in a selected area and support access to documents.

The Institution of Engineering and Technology - IET - developed a thesaurus for the Inspec database, which contains nearly 12 million bibliographic abstracts and indexing to journal articles, conference proceedings, technical reports and other literature in the fields of science and technology [4].

The Inspec Thesaurus contains about 18,400 terms, of which 9,573 are preferred terms, i.e. terms used in indexing, and 8,826 are lead-ins [5].

Term relationships show other terms in the thesaurus which are related to a given term. There are following types of relationship:

Hierarchical relationships:

NT narrower term
- from a given term to a more specific term

BT broader term
- from a given term to a more general term

TT top term
- from a given term to the most general term of the hierarchy

Associative relationships:

RT related term.
- This covers a range of non-hierarchical relationships, for example part-whole relationships.

Most relationships follow international standards - the TT relationship is a useful relationship beyond existing standards.

The IEEE Thesaurus is a controlled vocabulary of over 9,000 descriptive engineering, technical and scientific terms. Each descriptor included in the thesaurus represents a single concept or unit of thought. The descriptors are considered as preferred terms for use in describing IEEE content. [6]

II. STRUCTURES AND LINGUISTIC CONCEPTS OF THESAURI

A thesaurus consists of a partially ordered set of preferred terms. These terms are the basic semantic and cognitive units that are grouped together according to their meaning and reference fields. Terms represent objects or object types. Objects and the related terms can be characterized by attributes.

The theory for Formal Concept Analysis can be applied on thesauri [13].

Preferred terms in thesauri are concepts. A concept is based upon attributes and objects sharing common attributes [14]. For objects and attributes in each case a context has to be defined.

A ***contex***t *K* is a triple *K*(G,M,I) where G is a set of ***objects*** and M is a set of ***attributes***, and I ⊂ (G,M) is a relation that assigns objects g ∈ G to attributes m ∈ M.

For a subset A of objects of G, $A \subset G$, we define the ***derived subset*** A' of all attributes of the objects in A by

$A' = \{ a \in M \mid (g,a) \in I \text{ for all } g \in A \}$.

For a subset B of attributes of M, $B \subset M$, we define the ***derived subset*** B' of all objects matching with attributes in B by

$B' = \{ g \in G \mid (g,a) \in I \text{ for all } a \in B \}$.

C(A,B) is a ***concept*** of *K* , if

$$A \subseteq G$$
$$B \subseteq M$$
$$A' = B$$
$$B' = A.$$

Considering the set Φ of all concepts for the context K, an ordering relation can be introduced for the concept set in the following way:

$C_1 \leqslant C_2$ if $A_1 \subseteq A_2$,

where $C_1(A_1,B_1)$ and $C_2(A_2,B_2)$ are arbitrary concepts.

$(\Phi, \leqslant)$ is a lattice [14].

There are several software systems for the analysis and graphical representation of a concept lattice [15].

In concept lattices the ordering relation is a core concept that has been discussed often - also in the context of thesauri. Besides these relations it is important to go deeper in relations dealing with semantic structures, for example analogy.

Given a set of attributes M of a context *K*(G,M,I) and the set of attributes Ψ(A) and Ψ(B) of objects A and B,

$\Psi(A) = \{ a \in M \mid (g,a) \in I \text{ for all } g \in A \} \subset M$ and
$\Psi(B) = \{ a \in M \mid (g,a) \in I \text{ for all } g \in B \} \subset M$

Objects A and B are **analog** if and only if :

1. The set of common attributes Γ of A and B is not empty:

$$\Gamma(A,B) = \Psi(A) \cap \Psi(B) \neq \emptyset$$

2. The set of common attributes Γ is larger than the set of different attributes Δ :

$$\Delta(A, B) = (\Psi(A) - \Psi(B)) \cup (\Psi(B) - \Psi(A))$$
$$|\Delta(A, B)| \leqslant |\Gamma(A,B)|$$

The analogy between objects A and B is a symmetrical relation. Further conditions can be added for special kinds of analogies.

Fig.1 deals with the analogy between the terms objects "castle" and "château". Both kinds of objects are stationary - in contrast to a tent for example. The "castle" is characterized as "fortified" - the "château" not. There are common attributes and differentiating attributes.

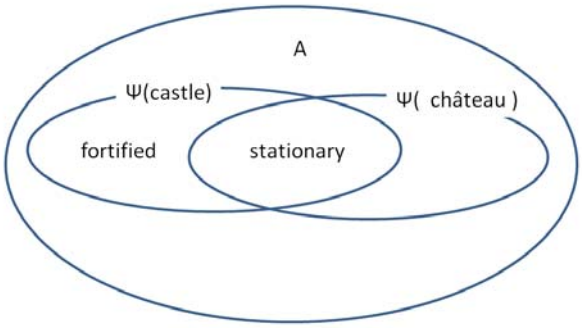

Figure 1. Analogy: Castle - château

There are many more relations to discuss from a semantic point of view, like antinomies or more specific concepts of analogy. A deeper understanding of these relations will help to use a thesaurus for stimulating creativity and creative problem solving.

III. TOPIC MAP BASED IMPLEMENTATION OF THESAURI

Topic Map based implementations of thesauri have been developed based on Ontopia software [16].

In our research in one case a topic map includes a partial thesaurus for IDAACS-like conferences based upon the IEEE-Thesaurus. Besides the terms and rations of the thesaurus this topic map includes documents of conferences as occurrences [17].

Topic Maps were formerly established as a standard for the representation and interchange of knowledge based upon ISO standard ISO/IEC 13250. In this perspective, a topic map for a thesaurus may describe:

- Topic types like “preferred terms”
- Association types like “broader terms”
- Topics of a given topic type like “project management”
- Associations a given topic type like “project management is a broader term with respect to project audit”
- Occurrences like papers of a conference on project management, where the topics and keywords of the conference for example are available as topics in the topic map.

The design and application of topic maps is supported by software – in the cases mentioned here we used Ontopia, which consists of:

The engine

The *heart* of Ontopia, since it takes care of storing and providing access to the topic maps in the system. Essentially, the engine is a set of Java APIs, for

- Accessing and modifying any part of a Topic Map
- Importing topic maps from file (in XTM 1.0, 2.0, CTM, TM/XML, or LTM format)
- Exporting topic maps to file (to XTM 1.0, 2.0, TM/XML, or LTM format)
- Converting from RDF to topic maps (and vice versa)

The *user interface* to enter all types of object types and objects for a topic map.

Omnigator

The web-based *topic map browser* which can display any topic map.

Vizigator

The *tool* showing graphical visualizations of the structure of a topic map

Based upon Ontopia 3 topic map applications in the context of thesauri were established 2010 and 2011:

- An application for conference management with topics like author, paper, reviewer, session, keywords, and the conference papers as occurrences – based upon the diploma thesis of Michael Meschede.[17]
- An application for project management with core terms of project management as topics, thesaurus relations as associations, and explanations of terms as occurrences based upon the master thesis of Marium Sharif Khan.
- A prototype for a linguistic thesaurus with typical terms as topics and thesaurus relations as associations based upon the master thesis of Sylvie Reusch. [21]

Fig.2 shows a part of topic map for conference management with a contribution to IDAACS 2011.

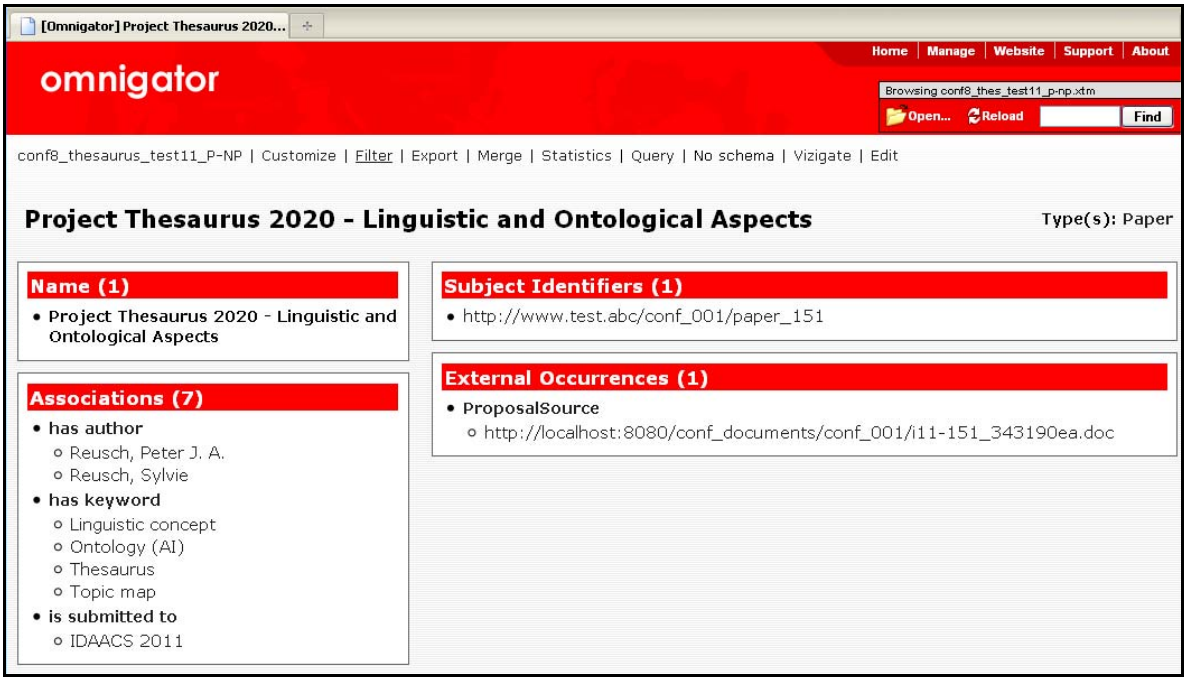

Figure 2. Topic map for conference management - here a paper with its associations

Fig. 3 shows that a conference contribution is related to several topics like author, paper, reviewer, session, keywords. These topics form a semantic web around the preferred term – here the conference paper. In the semantic web a conference contribution can be linked to other contribution, with help of associations like "share the same author" or "share the same keyword". [23]

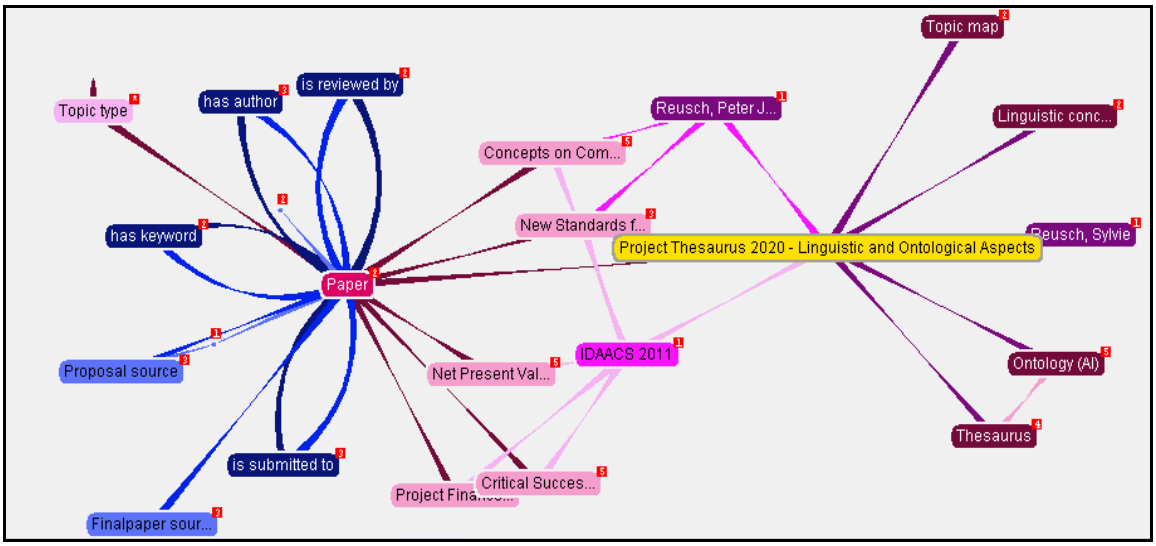

Figure 3. Topic relations for a specific conference contribution

In the next case there is a stronger focus on linguistic concepts and relations [18].

Fig.4 shows some topic types and association types of linguistic thesaurus. Here "preferred terms" are shown as the core node. Associations established links e.g. to "Related terms / RT objects". [22]

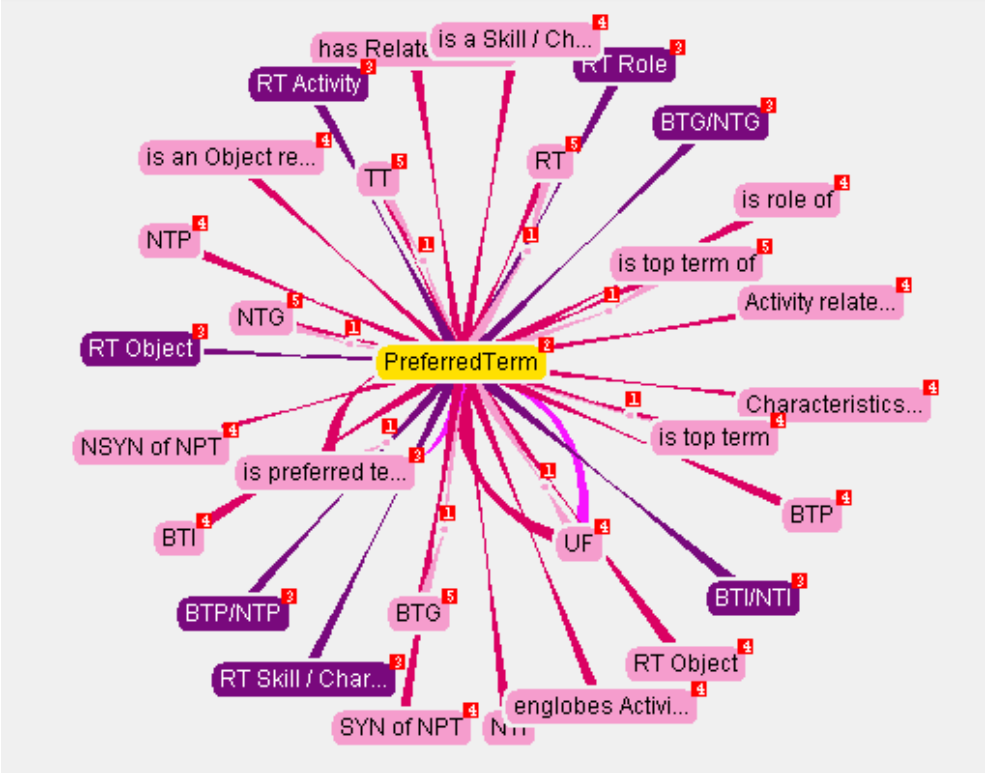

Figure 4. Topic types and association types for linguistic thesauri

IV. Results and Conclusions

In our research we analyzed core concepts of thesauri to learn how to develop a thesaurus that meets future requirements - how to develop the Thesaurus 2020 - for the years to come. Regarding linguistic thesauri the core structure of Roget´s thesaurus was identified as still strong.[21] Other thesauri like Thesaurus Larousse should come back to the basic approaches of Roget.[21] Mathematical concepts and IT concepts can support the development of thesauri.[21] The results of the applications based upon topic maps and Ontopia showed us that topic maps are appropriate for the implementation of thesauri and for applications that include a thesaurus and further elements like explanation of terms typical of glossaries or the papers of a conference.[17], [23] Topics, associations and occurrences establish an appropriate platform for such applications. Finally it was a core experience that we could easily merge parts of topics maps of all three applications. [23]

References

[1] Roget, Mark. Thesaurus of English words and Phrases. 150 th Anniversary Edition edited by George Davidson, London 2002.

[2] Robertson, Théodore. Dictionnaire idéologique. Recueil des mots, des phrases, des idiotismes et des proverbes de la langue française classés selon l'ordre des idées, Paris London 1859.

[3] Péchoin, Daniel. Thésaurus Larousse. Des idées aux mots et des mots aux idées, Paris 1992.

[4] Quemada, Bernard. Les Dictionnaires du Francais moderne 1539 - 1863. Etude sur leur histoire, leurs types et leurs méthodes, Paris 1967.

[5] Steger, Hugo (et al.). Dictionaries. An International Encyclopedia of Lexicography. New York 1989

[6] Bianucci, Gabriele (et.al.) 1992. "L'usage de relations sémantiques dans l'élaboration de thesauri: L'expérience du PTP (Petit Thesaurus Politique)," *in: Cahiers de lexicologie* LXI:2 , 59-84.

[7] Monneret, Philippe. Essais de linguistique analogique, Dijon 2004

[8] Péchoin, Daniel. Thésaurus documentaire, thésaurus idéologique et dictionnaire analogique. In:: Cotten, J.-P.;Hufschmitt B. Repérer, formaliser, traduire les concepts philosophiques, Colloque de Besancon 23-24 juin 1999, Paris 2001.

[9] Reusch, Sylvie: Thesaurus 2020, International Conference on Current Issues of Business and Society Development, University of Latvia, Riga 2011.

[10] The Institution of Engineering and Technology - IET - www.theiet.org

[11] IET Inspec Thesaurus 2010 - www.theiet.org

[12] IEEE Thesaurus 2009 - http://www.ieee.org/documents/2009Taxonomy_v101.pdf

[13] Ganter, Bernhard; Stumme, Gerd; Wille, Rudolf (eds.): Formal Concept Analysis: Foundations and Applications (Lecture Notes in Computer Science), Berlin 2009.

[14] Kovács, László : Concept Lattice Structure with Attribute Lattice, http://www.bmf.hu/conferences/HUCI2003/kovacs.pdf.

[15] http://www.fcahome.org.uk/

[16] www.ontopia.net

[17] Meschede, Michael: Ontologien und Topic Maps zur Unterstützung des Konferenzmanagement, Diplomarbeit Dortmund 2011.

[18] Reusch, Sylvie; Reusch, Peter: A Topic Map for Linguistical Thesauri, International Research Conference, Dortmund 2011.

[19] Casares, Julio: Diccionarioa Ideológico de la Lengua Español, Barcelona 1942.

[20] Sanders, Daniel: Deutscher Sprachschatz, Hamburg 1873-1877, reprint Tübungen 1985.

[21] Reusch, Sylvie: Des idées aux mots, des mots aux idées. Étude du thésaurus francais, Dijon, Mainz 2011.

[22] Reusch, Sylvie: The crisis of Thesauri today and concepts for future Thesauri, International Research Conference at the University of Applied Sciences and Arts in Dortmund, Dortmund 2011.

[23] Meschede, Michael; Reusch, Peter: Topic Maps for Thesauri - Concepts and Case Studies, International Research Conference at the University of Applied Sciences and Arts in Dortmund, Dortmund 2011.